American Fern Journal 99(2):78–92 (2009)

Molecular Cloning and Sequence Analysis of Cyanovirin-N Homology Gene in *Ceratopteris thalictroides*

XIAOQIONG QI
Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan, Hubei 430074, China
YONGXIA YANG
Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan, Hubei 430074, China
YINGJUAN SU*
School of Life Sciences, Sun Yat-sen University, Guangzhou, Guangdong 510275, China
TING WANG*
Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan, Hubei 430074, China

ABSTRACT.—A new full-length genomic DNA, encoding a member of the cyanovirin-N (CV-N) homologous protein family, has been cloned from the fern species *Ceratopteris thalictroides* by chromosome walking. It is 1993 bp long, contains a 723 bp open reading frame (ORF) that encodes a deduced protein (named CtCVNH) with 150 amino acid residues. CtCVNH has a predicted isoelectric point (PI) of 4.47 and a calculated molecular mass 15.9556 kDa. It possesses the conserved anti-HIV (human immunodeficiency virus) CV-N domain, which is the same as the cyanovirin-N homology (CVNH) members that were isolated from filamentous ascomycetes and *C. richardii*. Modeling of the tertiary structure indicated that CtCVNH is an elongated, largely β-sheet protein that displays internal two-fold pseudosymmetry. Comparative structure analysis of the predicted CtCVNH with native CV-N revealed that the major evolutionary changes occurring during the evolution of plant CVNHs were: 1) a length increase at N- and C-terminal regions; and 2) a loop to helix transition at the helical-turn regions. Phylogenetic analysis showed that CtCVNH was grouped together with the two CVNHs from *C. richardii*.

KEY WORDS.—*Ceratopteris thalictroides*, chromosome walking, single oligonucleotide nested PCR, inverse PCR, thermal asymmetric interlaced PCR, CVNH, bioinformatic analysis

Cyanovirin-N (CV-N) is an 11 kDa anti-HIV (human immunodeficiency virus) protein originally isolated from the extract of the cyanobacterium *Nostoc ellipsosporum* (Des.) Rabenh. (Boyd *et al.*, 1997). It consists of a single chain with 101 amino acids, exhibits significant internal sequence duplication between residues 1–50 and 51–101, and contains two intramolecular disulfide bonds (Gustafson *et al.*, 1997). CV-N is largely comprised of β-sheets with a two-fold pseudosymmetry (Bewley *et al.*, 1998). Its antiviral activity depends on the high-affinity binding to the HIV surface envelope glycoprotein, gp120 (Boyd *et al.*, 1997; Mori *et al.*, 1997). CV-N can specifically interact with high mannose groups (Bolmstedt *et al.*, 2001; Botos *et al.*, 2002), thereby blocking the interaction between gp120 and the receptor CD4 on target cells (O'Keefe *et*

* corresponding authors: Yingjuan Su, Tel: +86-20-84035090, Fax: +86-20-84036215, E-mail: suyj@mail.sysu.edu.cn; Ting Wang, Tel: +86-27-87510677, Fax: +86-27-87510251, E-mail: tingwang@wbgcas.cn.

al., 1997). Besides HIV strains (Boyd *et al.*, 1997), CV-N is also able to inactivate simian immunodeficiency virus (SIV), Ebola virus (EBO), herpes simplex virus-1 (HSV-1), and hepatitis C virus as well (Barrientos *et al.*, 2003; O'Keefe *et al.*, 2003; Barrientos and Gronenborn, 2005; Helle *et al.*, 2006). The potent inactivation of HIV plus unique biophysical properties make CV-N a candidate for a topical anti-HIV microbicide. The CV-N preclinical development is underway (Colleluoria *et al.*, 2005).

Recently, a family of CVNH (cyanovirin-N homology) has been identified. All CVNH proteins share a common fold that matches the one previously thought to be unique in CV-N (Percudani *et al.*, 2005). Current research on CVNHs is mainly focused on structural information, antiviral activity, carbohydrate-binding specificities or structure-function relationships (Percudani *et al.*, 2005; Koharudin *et al.*, 2008). For example, solution structures of three CVNHs from *Tuber borchii* Vittad., *Ceratopteris richardii* Brongn., and *Neurospora crassa* Shear et Dodge have been determined (Koharudin *et al.*, 2008) and may be helpful in elucidating the roles that these proteins play in the organs and during evolution.

CVNHs show a patchy organism distribution regarding the anti-HIV domain. They are present in organisms as diverse as cyanobacteria, filamentous ascomycetes and seedless plants (Percudani *et al.*, 2005). However, among plants, CVNHs have only been identified in the fern *C. richardii* until now. To provide useful information for understanding the evolution of CVNHs and developing antiviral polypeptides, here we report the cloning and sequence analysis of the full-length CVNH genomic DNA in *Ceratopteris thalictroides* (L.) Brongn. together with an analysis of CVNHs phylogeny and modeling of the protein tertiary structure.

Materials and Methods

Plant materials.—*Ceratopteris thalictroides* was collected from Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan, Hubei, China. Young and healthy leaves were sampled, immediately frozen in liquid N_2, and stored at −70°C until used.

Genomic DNA extraction.—Total genomic DNA was extracted from fresh leaves following the modified CTAB protocols (Su *et al.*, 1998). DNA concentration and purity were determined by measuring UV absorption using a Pharmacia 2000 UV/Visible spectrophotometer. DNA intactness was checked by 1.0% agarose gel electrophoresis.

Molecular cloning of the full-length genomic DNA.—Based on the *C. richardii* EST sequence (Accession No. BQ087187), specific primers were designed to amplify the internal region of *CVNH* in *C. thalictroides*. The forward primer CVNH-F was 5′-GTGGGCGTCTAGCGATTTCCTTT-3′, and the reverse primer CVNH-R was 5′-ATCATCCGCTGCTTGCTTCTTCG-3′. The reaction mixture (20 μL) contained 50 ng template DNA, 40 pmol each primer, 1 pmol each dNTP, 1.0 U *Taq* DNA polymerase and 1 × *Taq* polymerase buffer. PCR was performed using the following protocol: the template was

TABLE 1. The primers used in chromosome walking.

Primer name	Primer sequence
5′IPCR	5′-GGTGATATTGCCCGTCGGTGCCTTT-3′
5′SON-1	5′-ATCACTGTTGAGGCAATCTGCGGCT-3′
5′SON-2	5′-GCTGCGATCAAGACGATGAGAAAAC-3′
5′SON-3	5′-CCATCGCTTCTAGGAGTAAACAGAC-3′
3′TAIL-1	5′-GTGCAAAGGCACCGACGGGCAATAT-3′
3′TAIL-2	5′-GGGGTGTTGGATTTCTGTGGCTATG-3′
3′TAIL-3	5′-AAGCGAAGAAGCAAGCAGCGGATGA-3′

denatured at 94°C for 5 min followed by 36 cycles of amplification (94°C for 50 s, 61°C for 50 s, 72°C for 90 s) and a final extension of 10 min at 72°C.

Based on the sequence obtained from the internal DNA region, two sets of nested primers for 5′ single oligonucleotide nested PCR (SON-PCR) (Antal *et al.*, 2004) and 3′ inverse PCR (IPCR) (Triglia *et al.*, 1988) combined thermal asymmetric interlaced PCR (TAIL-PCR) (Liu and Whittier, 1995) were designed to amplify the 5′ and 3′ flanking sequences. These primers included 5′IPCR, 5′SON-1, 5′SON-2, 5′SON-3, 3′TAIL-1, 3′TAIL-2, and 3′TAIL-3 (Table 1, Fig. 1). They were of high annealing temperatures and synthesized by Invitrogen (Shanghai).

The 5′ flanking sequence was amplified by SON-PCR. The primary PCR was carried out in a 20 μl volume containing 50 ng genomic DNA, 50 pmol single primer (5′SON-1), 50 mol/L each dNTP, 2.0 U *Taq* DNA polymerase and 1 × *Taq* polymerase buffer. For the secondary PCR, two single primers (5′SON-2 and 5′SON-3) were separately used. The reaction solution was the same as that of primary PCR except that 1 μl of a 1:50 dilution of the primary PCR products was used as the template.

The 3′ flanking sequence was obtained using IPCR combined TAIL-PCR. *Ceratopteris thalictroides* genomic DNA was digested with Pac I (NEB, BSA 5 U μg^{-1} of DNA) at 37°C for 3 h, and then heated at 65°C for 20 min. The digested DNA was self-ligated overnight at 15°C with a concentration of 0.3–0.5 μg/ml in the presence of 3 U/ml T4 DNA ligase (Promega). PCR was carried out in a 20 μl volume with 1 μl ligated product, 1 pmol each dNTP, 40 pmol each primer (5′IPCR and 3′TAIL-1), 1.0 U *Taq* DNA polymerase and 1 × *Taq* polymerase buffer. The primary PCR of TAIL-PCR was performed using primer

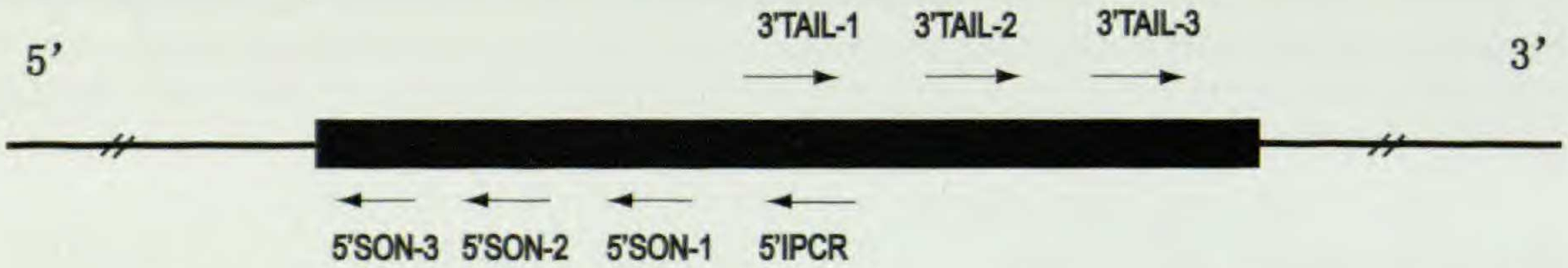

FIG. 1. Schematic view of position and orientation of nested primers used in this study and of their relative positions to the amplified sequence of the specific PCR. The rectangle frame indicates the sequence obtained by specific PCR, whereas the line represents regions determined by further chromosome walking.

TABLE 2. Cycling conditions used for SON-PCR, IPCR, and TAIL-PCR.

Name	Reaction	Cycle no.	Thermal condition
5′SON-PCR	Primary	1	94 °C (5 min)
		5	94 °C (30 s), 65 °C (1 min), 72 °C (2.5 min)
		1	94°C(30 s), 29°C(3 min), ramping to 72°C over 3 min, 72°C(2.5 min)
		60	94°C(30 s), 65°C(1 min), 72°C(2.5 min)
		1	72°C(7 min)
	Secondary	1	94°C(5 min)
		30	94°C(30 s), 65°C(1 min), 72°C(2.5 min)
		1	72°C(7 min)
3′IPCR	IPCR	1	94°C(10 min)
		33	94°C(1 min), 65°C(1 min), 72°C(2 min)
		1	72°C(10 min)
3′TAIL-PCR	Primary	1	93°C(1 min), 95°C(1 min)
		5	94°C(30 s), 65°C(1 min), 72°C(2.5 min)
		1	94°C(30 s), 25°C(3 min), ramping to 72°C over 3 min, 72°C(2.5 min)
		15	94°C(30 s), 65°C(1 min), 72°C(2.5 min)
			94°C(30 s), 65°C(1 min), 72°C(2.5 min)
			94°C(30 s), 44°C(1 min), 72°C(2.5 min)
		1	72°C(5 min)
	Secondary	1	94°C(1 min)
		15	94°C(30 s), 65°C(1 min), 72°C(2.5 min)
			94°C(30 s), 65°C(1 min), 72°C(2.5 min)
			94°C(30 s), 44°C(1 min), 72°C(2.5 min)
		1	72°C(5 min)

3′TAIL-1 as the gene-specific primer and primer AD (5′-TC(G/C)TICGNA-CIT(A/T)GGA-3′) (Liu and Whittier, 1995) as the arbitrary degenerate primer in a total 20 µl volume that contained 1 µl of a 1:50 dilution of the IPCR products, 2 pmol each dNTP, 40 pmol primer 3′TAIL-1, 500 pmol primer AD, 2.0 U *Taq* DNA polymerase and 1 × *Taq* polymerase buffer. For the secondary PCR, two gene-specific primers (3′TAIL-2 and 3′TAIL-3) were separately used with the same arbitrary primer as used in the primary one. The reaction solution was the same as that used for the primary PCR except that 1 µl of a 1:50 dilution of the primary PCR products was used as the template. Thermocycling profiles used for SON-PCR, IPCR, and TAIL-PCR are listed in Table 2.

Recovery of PCR products.—PCR products were purified by running them through a 1.0% low melting agarose gel. The desired DNA band was cut out and recovered using the DNA rapid purification kit (Omega).

DNA cloning and sequencing.—A purified PCR product was ligated into a pMD 19-T (TaKaRa) vector and then used to transform competent *Escherichia coli* cells DH-5α. A positive clone was identified by blue/white selection and ascertained by PCR. Purified plasmid DNA was sequenced in both directions by standard methods on an ABI 3730 automated sequencer at Invitrogen (Shanghai). Primers M13F and M13R located on pMD19-T vector were utilized for sequence determination.

In silico *analysis and molecular modeling*.—ORF finder was used to predict coding sequence, and promoter analysis was performed online (http://www.fruitfly.org/cgi-bin/seq_tools/promoter.pl). Sequence analysis was conducted using the BLAST program (Altschul *et al.*, 1997) and other programs available at the ExPASy server (Gasteiger *et al.*, 2003). Multiple sequence alignment was carried out using the ClustalX software (Thompson *et al.*, 1997). Figures of multiple sequence alignment adorned with secondary structure elements were generated with ESPript (Gouet *et al.*, 1999). Primary structure analysis of the deduced CtCVNH (CVNH protein from *C. thalictroides*) was conducted with ProtParam (Gasteiger *et al.*, 2005) by using the ExPASy server online (http://www.expasy.ch/tools/protparam.html). Secondary structure was predicted with SOPMA program (Geourjon and Deleage, 1995) online (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_sopma.html). Phylogenetic analysis was carried out using programs from the PHYLIP package; genetic distances were estimated with PROTDIST using the Jones-Taylor-Thornton model of amino acid substitutions. Neighbor-joining trees (Saitou and Nei, 1987) were constructed using the NEIGHBOR program; 1000 random replications were utilized for bootstrap analysis, which was performed with the SEQBOOT and CONSENSE programs. Phylogenetic trees were rendered with the TREEVIEW program (Page, 1996). The three-dimensional (3D) structural models of CtCVNH were built by the homology-based method using the SWISS-MODEL program (Guex and Peitsch, 1997; Schwede *et al.*, 2003; Arnold *et al.*, 2006). The template used for modeling was *C. richardii* CVNH (PDB code 2jzjA) (Koharudin *et al.*, 2008). Models were displayed with the PyMol program (Delano, 2002).

Results and Discussion

Molecular cloning of the full-length genomic DNA.—Using a pair of specific primers (CVNH-F and CVNH-R), a single fragment of 775 bp was amplified from the *C. thalictroides* DNA [Fig. 2(a)]. Compared with the *C. richardii* cDNA sequence (Accession No. BQ087187), the sequence from *C. thalictroides* has two additional fragments that do not exist in *C. richardii* cDNA and the remaining parts of the sequence are identical to the *C. richardii* cDNA (Fig. 3). The *CtCVNH* intron–exon boundaries were thus deduced; it is composed of three exons and two introns. Based on the amplified sequence of the specific PCR, two sets of nested primers were further designed to obtain the 5′ and 3′ flanking sequences, respectively. A clear single band ≈ 800 bp of the 5′ flanking sequence was generated in the secondary reaction [Fig. 2(b)] using SON-PCR, while a ≈ 750 bp 3′flanking sequence was amplified through IPCR combined TAIL-PCR [Fig. 2(c)].

Sequence analysis of the CtCVNH *gene*.—The cloned full-length *CtCVNH* gene is 1993 bp in length, including a 818 and 452 bp 5′ and 3′ untranslated region (UTR) respectively, and a 723 bp coding region. The 5′UTR has a TATA box in the predicted promoter elements. The ATG start codon, which is numbered +1 to +3, is flanked by G in both positions -3 (3 nucleotides before

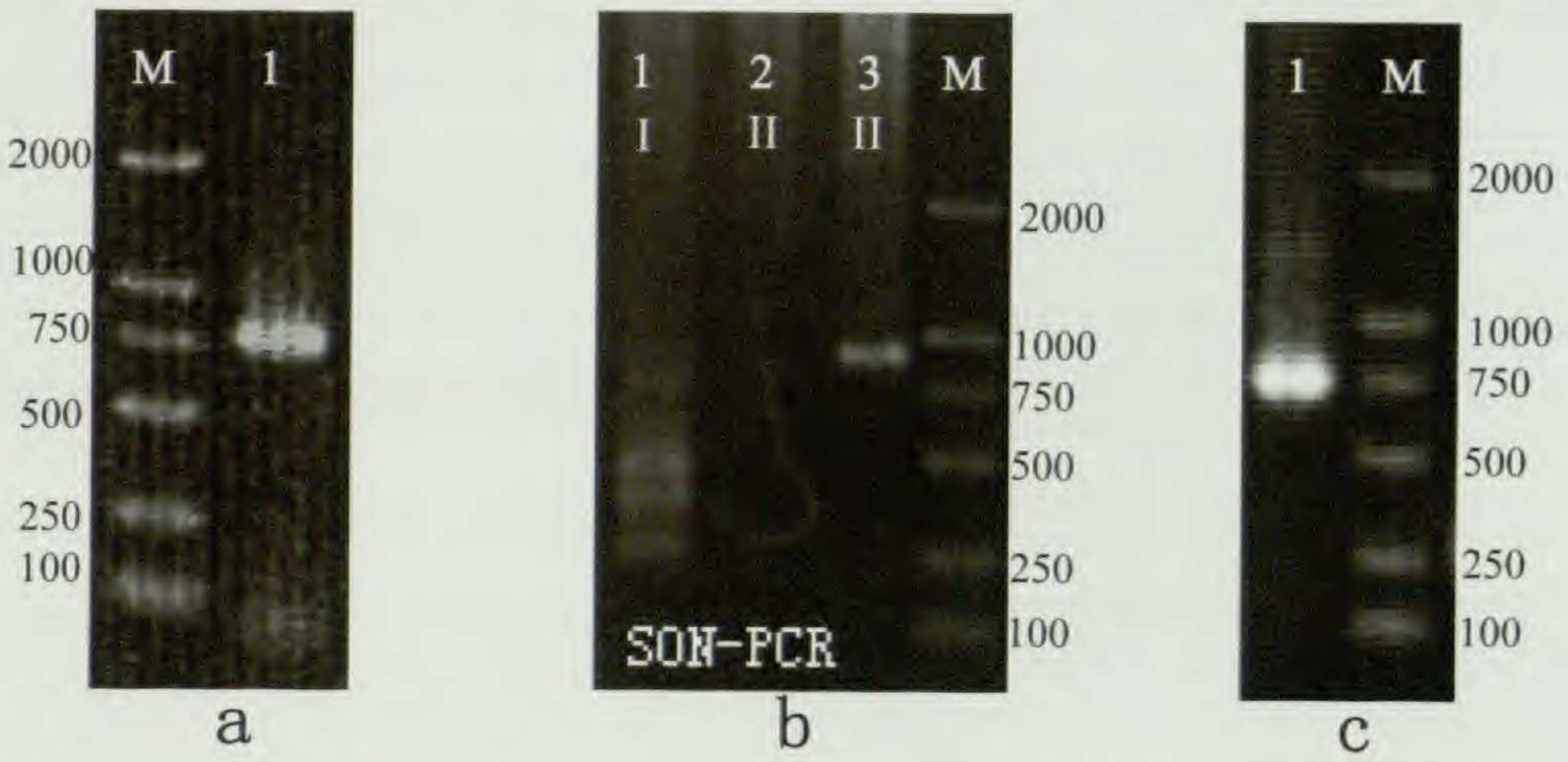

FIG. 2. Agarose gel electrophoresis of the specific PCR(a), SON-PCR(b), and IPCR+TAIL-PCR(c) products. M is molecular weight marker (DL2000). a 1: about 750 bp fragment generated by specific PCR with primer CVNH-F and CVNH-R. b 1: smear bands produced by the first reaction of SON-PCR with primer 5′SON-1. 2: no clear band produced by the secondary reaction of SON-PCR with primer 5′SON-2. 3: a single band obtained by the secondary reaction of SON-PCR with primer 5′SON-3. c 1: the amplified DNA fragment of secondary reaction of TAIL-PCR with primer 3′TAIL-2 and AD.

the ATG codon) and +4 (1 nucleotide after the ATG codon), indicating that it is located in a sequence context for strong translational initiation (Kozak, 1999). The 3′UTR has a polyadenylation signal (AATAAA) and six ATTT domains (Fig. 4). These ATTT domains may be important for mRNA destabilization (Shaw and Kamen, 1986). The *CtCVNH* gene encodes a deduced protein of 150 amino acid residues with a predicted isoelectric point (pI) of 4.47 and a calculated molecular mass of 15.9556 kDa. Regarding its amino acid composition, the most abundant is Ser (13.3% by frequency), followed by Gly (9.3%), Leu (9.3%), Ala (7.3%), Asn (7.3%), Asp (7.3%), and Val (6.7%). Acidic and basic amino acids constitute 10.0% and 5.3% of the protein, respectively. Moreover, 15.3% of the amino acids are charged, and the percentages of polar and hydrophobic amino acids are 64% and 25.3%, respectively (Table 3).

With regard to the predicted secondary structure, the CtCVNH protein consists of 16.00% alpha helices, 28.67% extended strands, 12.67% β turns, and 42.67% random coils. The extended strands and random coils constitute the interlaced domain of the main part of the secondary structure.

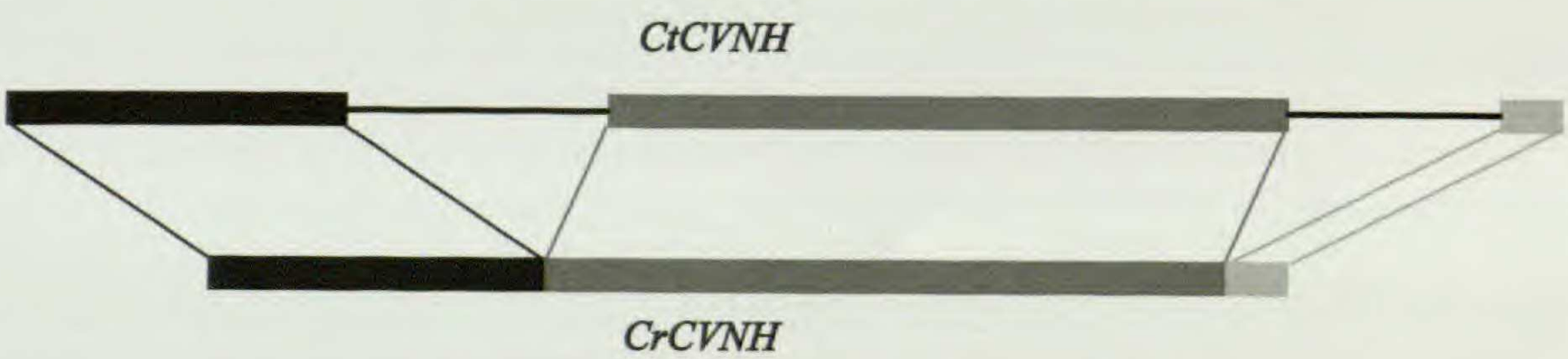

FIG. 3. Schematic view of the exon and intron positions deduced from *C. richardii* cDNA. The exon and intron are indicated by rectangle frame and line, respectively.

```
                                        CCATCGCTTCTAGGAGTAAACAGACATACTATAACTGC   -781
ACTCTAACACATTTTCAACAATACAAGAAACATTTATACATACTTCACACGTGCACATCA   -721
TCACACATACACACATACTATGTGTGCATACATGCTCACATGCATAGTAGCATGCATATA   -661
CACATAGTAAATAATGCACATGCATACATATATACACATGCACATTATTTCGGTGGGCCT   -601
TCAAATGCTAGTTAAAAATTTATTGTTATTATCTTTTTAAATGGAGTTCCGGCAAGTGGT   -541
TACTCCCCCGAGCTTTAACTTTGAAAGAATAACTTAACACGTGAATGGACATAGTGTTCG   -481
AGGGACCCCTCGTATCCCACAAAGTAAGAGTTGGATAAAATTACAACTTGTTTCCATTGG   -421
TTTAGACTGGCCAAGGCCATTACACGCTCAAGCCAGAGATTGAACTTTGAACTTCCACCT   -361
GAAGAGAGCAAAGACTCAACCACTAGAACAACGAAATGTGAACTTAAAAATTTATTATTG   -301
GTTGTTTTGAAAAAAATTCTTTATTACAATCTTAATTTTCATTTGATTGCCCAATTTTTA   -241
GAAAGGCAGTCACGGAGACCAAACGCTAAGTCCACTGCTCGTTTGCAAAGGAGAACGTTT   -181
GTCTTTATCGAGTATTATATACAGAAACGCTAAGTTGCGCCACATGTCATGAAATAAGCG   -121
TCCGAGATCCAAAGACGTTAACTTTCCCTACTATATAAATGGGGAAGCTTGAGACCATCT    -61
GCCCAGTCCGTTATTCGTCTAGCGATTTCCTTTCGAAGTGTCTGTTTACTCCTAGAAGCG     -1
ATGGAGTACTTTACACCGCGATATCTTCTCCTTCAAGGTTTTCTCATCGTCTTGATCGCA     60
M  E  Y  F  T  P  R  Y  L  L  L  Q  G  F  L  I  V  L  I  A
GCATCGAATGCCAGCGCTCAAgtaagtctttctgatttccagtttcatatgttaaatctc    120
A  S  N  A  S  A  Q
tctctctctctctctctctctctctctctctcccaattctggtttgtgatgttgtatc     180
atcgttctcgtctgcaTGCGATTTCTCATACTCGTGCAAGGATGTGACTGTAACCGGCAA    240
                  C  D  F  S  Y  S  C  K  D  V  T  V  T  G  N
CTTCCTAGCCGCAGATTGCCTCAACAGTGATGGTGCATACGATCGGTCTTCTCTGAATAT    300
 F  L  A  A  D  C  L  N  S  D  G  A  Y  D  R  S  S  L  N  M
GAACGACATGATTGGTAATAGTAATGGAAGGCTTGTATTTCCCGGTACCTCCTTCCGTAA    360
 N  D  M  I  G  N  S  N  G  R  L  V  F  P  G  T  S  F  R  N
TTCATGCTTGAGTGTGGAGATCAACGACGGTCATACGCTCACAGCTTCGTGCAAAGGCAC    420
 S  C  L  S  V  E  I  N  D  G  H  T  L  T  A  S  C  K  G  T
CGACGGGCAATATCACCCTACAAGCCTTGATCTCAACTCTTGCGTTTATAATGCTGACGG    480
 D  G  Q  Y  H  P  T  S  L  D  L  N  S  C  V  Y  N  A  D  G
GGTGTTGGATTTCTGTGGCTATGGTGTCGGAAAATCAACAGCCTACGTCAAGTCCAGTAC    540
 V  L  D  F  C  G  Y  G  V  G  K  S  T  A  Y  V  K  S  S  T
CGTgtaatgtccctgcaacaagtactgccgtagttatattatatatcgttacttcacct     600
 V
ctcaggaaatgctttactttgcgctctacaacactagctattctacttatatatggcatt    660
tgctggacgttgtttatattaattttctttctcttcacAAGCGAAGAAGCAAGCAGCGGA    720
                                      S  E  E  A  S  S  G
TGATTTGAGCATGTGCTTCACTGTCTGCAAACTTCCTCTATTATGCTGTGACGTTGTCTA    780
*
TCGAAGCCACACATCGGATTATATATAACATAGCGTCTTTTCATTTGGGTGATATGCTCC    840
TCGCGTTCTCTTTCCCTGCTCTTTCTTTGTCTTTGTTATGCCGCCTCAAGTTCTGCATGT    900
ATGATTTTATTCCTACGTGCTGTATTGATATCAGTAGATGTGTCCTATATTCATTTCACC    960
CTTCATTAATTAAGTATGTCAATGTCAACAGAAGTTACTCTCTAATAAAATCCATACGAA   1020
ATCTCCTGCATTCAGAAACCCACAAATAGAGAATTTCTAACAATGTCAGTTTATTACGTC   1080
TATTTATTCATAACCCACATCTAACTAACGCGAGTTACACTATAGGAAATCTCTTTCTAC   1140
TATTTTGGATTGTTAGTTTATTACGTCTATTTATT                            1175
```

FIG. 4. Nucleotide and deduced protein sequences of *CtCVNH* gene. The predicted amino acid sequence is shown below its open reading frame. The predicted promoter sequence is shown in shaded box (transcription start site shown in larger font). The TATA box is boxed with solid lines. The polyadenylation signal is underlined and boldface. The ATTT regions of the 3′UTR are underlined. The introns are present in lowercase letters.

TABLE 3. Component analysis of amino acid sequence of CtCVNH.

Amino acid	Number	Frequency(%)
Hydrophobic amino acid	38	25.3
Charged amino acid	23	15.3
Polar amino acid	96	64.0
Acidic amino acid	15	10.0
Basic amino acid	8	5.3
Ala(A)	11	7.3
Gly(G)	14	9.3
Met(M)	3	2.0
Ser(S)	20	13.3
Cys(C)	7	4.7
His(H)	2	1.3
Asn(N)	11	7.3
Thr(T)	10	6.7
Asp(D)	11	7.3
Ile(I)	4	2.7
Pro(P)	3	2.0
Val(V)	10	6.7
Glu(E)	4	2.7
Lys(K)	4	2.7
Gln(Q)	3	2.0
Trp(W)	0	0.0
Phe(F)	7	4.7
Leu(L)	14	9.3
Tyr(Y)	8	5.3
Arg(R)	4	2.7

Amino acid sequence alignment and phylogenetic analysis.—Initial homology searches were conducted with the deduced CtCVNH amino acid sequence in the non-human, non-mouse EST database at the NCBI (National Center for Biotechnology Information, NIH, Bethesda) by using the tblastn program (Altschul *et al.*, 1997). A new member of CVNHs was uncovered from the plant *Selaginella moellendorffii* Hieron. by conducting these searches. The results (Table 4) showed that the CVNH members were present in fungi and plants (E value < 0.01). Above 70% of the members occurred in fungi. A comparison of the deduced CtCVNH against other CVNHs revealed that CtCVNH shares a high degree of similarity with the two CVNHs from *C. richardii* (99% and 53% identity, respectively), and a reduced level of similarity with the CVNHs from fungi (26–33%). Multiple sequence alignment indicated that the anti-HIV domain is conserved [Fig. 5(a)]. The most conservative sites were F4, L18, G27, L36, G41, N42, G45, F54, L69, G78, L87, N93, and G96 (the numbering is in line with the *N. ellipsosporum* CV-N). These residues are predominantly located in the hydrophobic core region, which are involved in hydrophobic interactions between the β-hairpin and the underlying triple-stranded β-sheet of each repeat (Percudani *et al.*, 2005). Also conserved are hydrophilic amino acids involved in the formation of the hydrogen-bonded bridges that connect

TABLE 4. Members of the CVNHs identified in nucleic acid databases.

Accession[a]	Source	Amino acid	Organism	Classification	Percent ID[b]	E[c]
BQ087187	mRNA	150	*Ceratopteris richardii*	Filicales	99	2e-81
BQ087517	mRNA	142	*Ceratopteris richardii*	Filicales	53	3e-24
FE488500	mRNA	106	*Selaginella moellendorffii*	Selaginellales	31	0.002
CO138997	mRNA	106	*Aspergillus flavus*	Eurotiomycetes	30	1e-05
EY428974	mRNA	106	*Aspergillus oryzae*	Eurotiomycetes	30	1e-05
EH395121	mRNA	106	*Phaeosphaeria nodorum* SN15	Dothideomycetes	33	2e-04
AJ917737	mRNA	101	*Trichoderma stromaticum*	Sordariomycetes	30	9e-05
DR044956	mRNA	106	*Triticum aestivum /Phaeosphaeria nodorum* mixed EST library	-	33	1e-04
BQ110230	mRNA	143	*Verticillium dahliae*	Sordariomycetes	26	0.002
AJ915102	mRNA	105	*Hypocrea lixii*	Sordariomycetes	33	0.003
FG382361	mRNA	104	*Hypocrea virens*	Sordariomycetes	33	0.004

[a]For protein sequences deduced from mRNA data, the accession number of a representative EST sequence is reported.
[b]Percent identity with respect to the deduced CtCVNH.
[c]Expected values for pairwise comparisons (blosum62 matrix) based on the size of the non-human, non-mouse EST database.

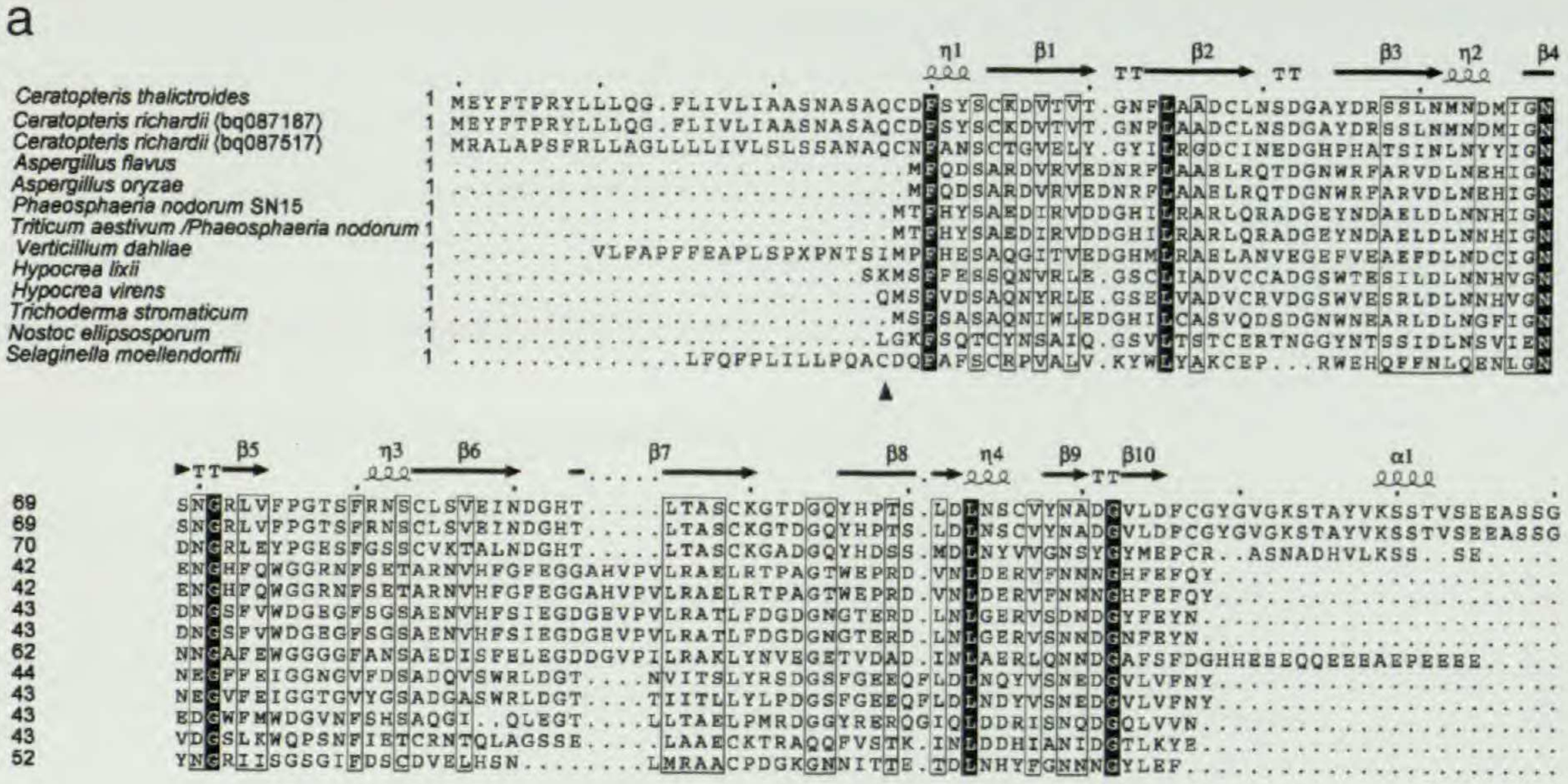

FIG. 5(a). Multiple alignment of the CVNHs. Sequence conservation is visualized according to the ESPript chemical equivalence measure, with a similarity threshold set to emphasize strictly conserved positions. Invariant residues are boxed in black; physico-chemical equivalent residues are boxed in gray. The deduced secondary structure of CtCVNH is shown above the alignment. The predicted N-termini of the mature fern polypeptides are indicated by a black triangle. Fig. 5(b). Comparison of the amino acid sequences of domains 1–76 and 77–150 of CtCVNH. Sequence homology of the domains was maximized by insertion of gaps (-). Identical amino acids (·) and conserved amino acids (*) are indicated.

β-strands 1–9 and 4–6 (Bewley *et al.*, 1998). These suggest a critical structural role or their involvement in carbohydrate binding.

Sequence similarity was also examined between the first (residues 1–50, according to the numbering in the *N. ellipsosporum* CV-N) and the second half (residues 51–101) of the CVNHs. Like CV-N, all CVNHs comprise two tandem sequence repeats with identities ranging from 24.0% to 41.1% (data not shown). Several residues are completely conserved [Fig. 5(b)]. The apparent sequence similarity between the two repeats (with an average identity of 33.3%) can be ascribed to the structural constraints imposed by the symmetrically interconnected CVNH fold (Percudani *et al.*, 2005).

A neighbor-joining tree (Fig. 6) was constructed to analyze the phylogenetic relationships of CtCVNH with other CVNHs (Table 4). It shows that CtCVNH is closely related to the member from *C. richardii* (BQ087187), and CVNHs

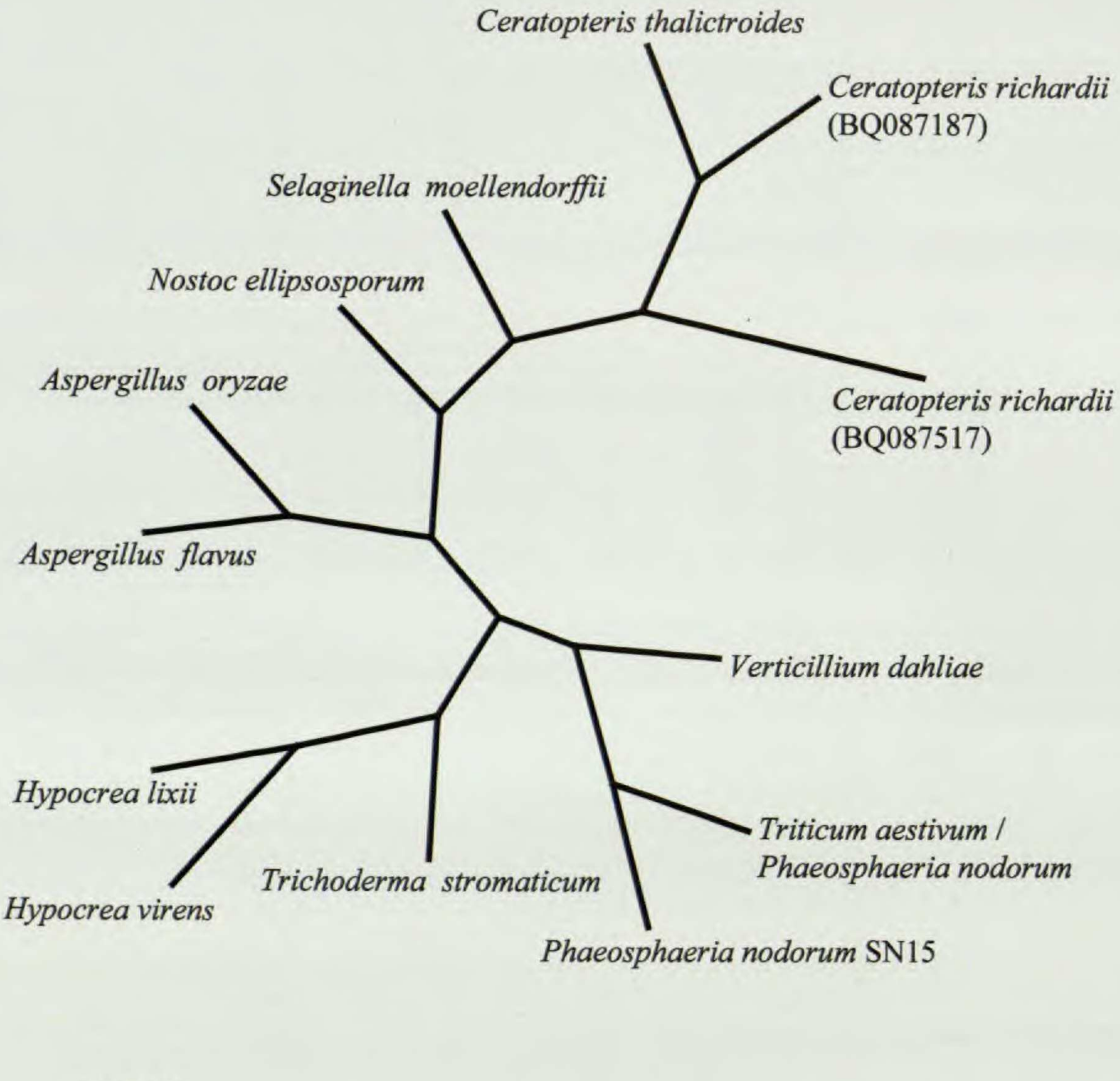

FIG. 6. Phylogeny of the CVNH proteins. The unrooted tree was constructed by neighbor-joining analysis (Saitou and Nei, 1987) of genetic distances estimated with the Jones-Taylor-Thornton model. Branch lengths are proportional to genetic distances as indicated by the scale bar representing 10 PAMs (point-accepted mutations).

belonging to different phyla form monophyletic groups. The CVNH domains may have common origin; however, Percudani et al. (2005) suggested that in fungi and seedless plants the domain has been separately amplified with different copy numbers following the separation of these two lineages.

Predicted CtCVNH tertiary structure and the structural evolution of CVNHs/CV-N.—Understanding the structural properties of CtCVNH is important for clarifying the conservation and variation of CVNHs as well as the roles they play in plants. *In silico* methods exist to predict with high reliability the tertiary structure of proteins from template structures (Saenz-Rivera *et al.*, 2004; Gopalasubramaniam *et al.*, 2008). Predicting a structure can yield insights into potential evolutionary patterns for CVNHs. Because CtCVNH and

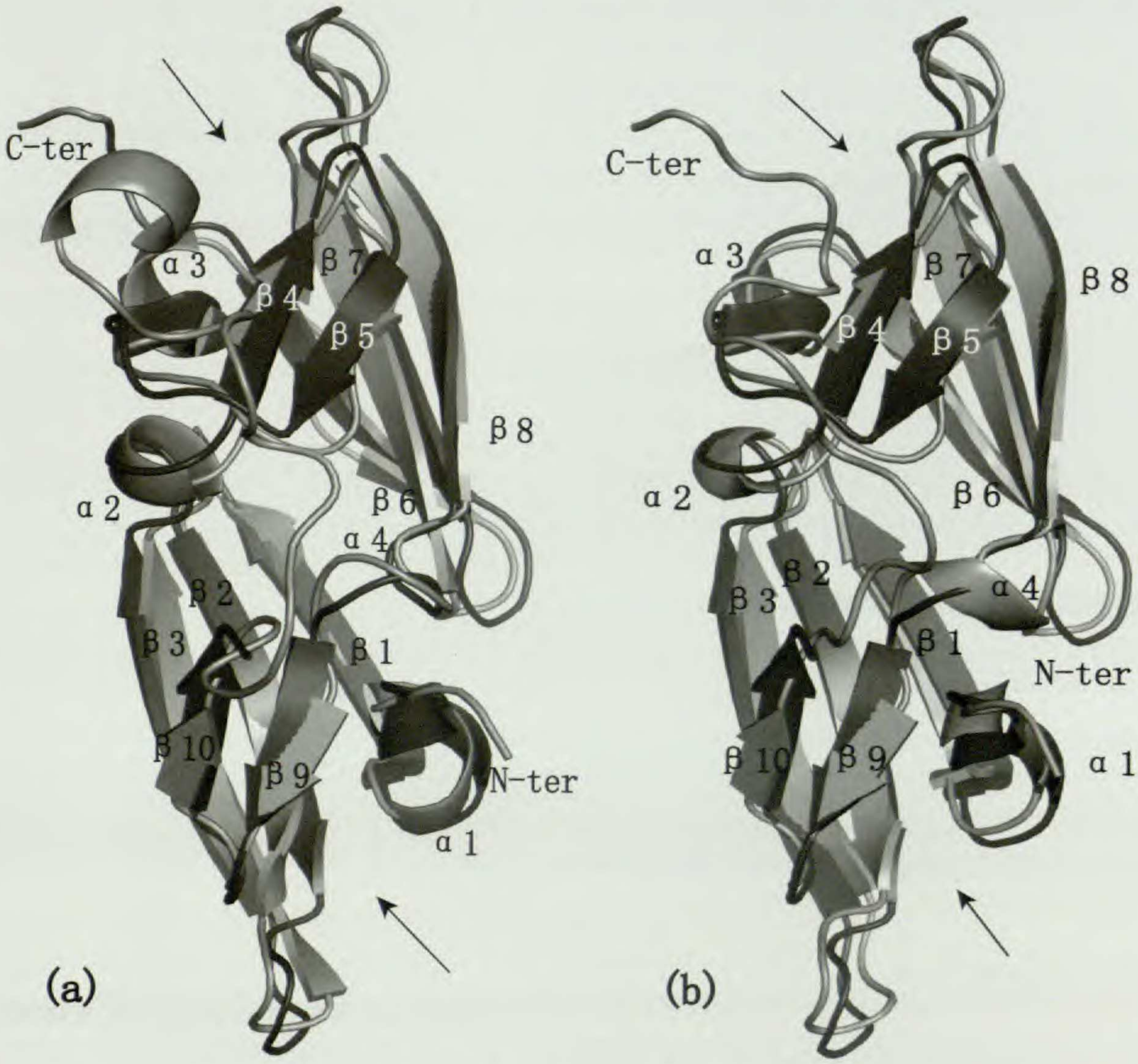

FIG. 7. Panel (a) overlay of predicted CtCVNH (gray) and native CV-N (black) tertiary structures. Panel (b) overlay of CrCVNH (gray) and native CV-N (black) tertiary structures. β-strands are indicated with β1–10 and helical turns with α1–4. Arrow shows the two sugar binding pockets of CV-N.

C. richardii CVNH (CrCVNH) are approximately 50% identical, we predicted the tertiary structure of CtCVNH using CrCVNH as a template. Fig. 7a further shows that the predicted CtCVNH comprises two tandem sequence repeats. They form equivalent, elongated structures via the combination of a triple-stranded β-sheet and a β-hairpin. Thus two symmetrically related fold-domains are created, each containing a sugar-binding site. Fig. 7a indicates that CtCVNH structure is quite similar to that of native CV-N, including the positions of triple-stranded antiparallel β-sheet (the first sequence repeat: β1, β2, and β3; the second: β6, β7, and β8), β-hairpin (formed by β4 and β5, β9 and β10, respectively), and α-helical turn (α1–4). However, the structures differ in that the N- and C-terminal regions are longer in CtCVNH than in CV-N, the helical turn (α3) folds differently, and an (3/4 turn) α-helix exists within the C-terminal region of predicted CtCVNH. Moreover, the β1 and β6 strands are

shorter in CtCVNH than in CV-N. To further understand the CVNH evolution in plants, we also compared the tertiary structure of CrCVNH with CV-N. Fig. 7b shows that the native CrCVNH structure is more similar to that of native CV-N, and most differences exist in the helical turn regions (α2, α3, α4) rather than in the β-strand ones. It is worthwhile to note that these differences are located in the sugar binding pockets of the proteins, which imply that CrCVNH and CV-N may have different affinities for mannose disaccharide ligands (Percudani *et al.*, 2005).

In conclusion, molecular cloning and characterization of CtCVNH showed that CtCVNH is very similar to other CVNHs from ascomycete fungi and the fern *C. richardii*, having a typical anti-HIV domain [Fig. 5(a), 7], indicating that CtCVNH belongs to CVNH family. This is the first time a full-length genomic DNA of CVNH in plants has been cloned. Our results provide a basis for a deeper understanding of CVNH function and evolution.

ACKNOWLEDGMENTS

This project was supported by the "100 Talent Project" of Chinese Academy of Sciences (Grant No.: 0729281F02), the National Natural Science Foundation of China (Grant No.: 30771763, 30170101), and the "Outstanding Young Scientist Project" of the Natural Science Foundation of Hubei Province, China (Grant No.: 0631061H01).

LITERATURE CITED

ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHAFFER, J. ZHANG, Z. ZHANG, W. MILLER and D. J. LIPMAN. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. 25:3389–3402.

ANTAL, Z., C. RASCLE, M. FEVRE and C. BRUEL. 2004. Single oligonucleotide nested PCR: a rapid method for the isolation of genes and their flanking regions from expressed sequence tags. Curr. Genetics 46:240–246.

ARNOLD, K., L. BORDOLI, J. KOPP and T. SCHWEDE. 2006. The SWISS-MODEL workspace: a web-based environment for protein structure homology modeling. Bioinformatics 22:195–201.

BARRIENTOS, L. G. and A. M. GRONENBORN. 2005. The highly specific carbohydrate-binding protein cyanovirin-N: structure, anti-HIV/Ebola activity and possibilities for therapy. Mini Rev. Med. Chem. 5:21–31.

BARRIENTOS, L. G., B. R. O'KEEFE, M. BRAY, A. SANCHEZ, A. M. GRONENBORN and M. R. BOYD. 2003. Cyanovirin-N binds to the viral surface glycoprotein, GP1,2 and inhibits infectivity of Ebola virus. Antivir. Res. 58:47–56.

BEWLEY, C. A., K. R. GUSTAFSON, M. R. BOYD, D. G. COVELL, A. BAX, G. M. CLORE and A. M. GRONENBORN. 1998. Solution structure of cyanovirin-N, a potent HIV-inactivating protein. Nat. Struct. Biol. 5:571–578.

BOLMSTEDT, A. J., B. R. O'KEEFE, S. R. SHENOY, J. B. MCMAHON and M. R. BOYD. 2001. Cyanovirin-N defines a new class of antiviral agent targeting N-linked, high-mannose glycans in an oligosaccharide-specific manner. Mol. Pharmacol. 59:949–954.

BOTOS, I., T. MORI, L. K. CARTNER, M. R. BOYD and A. WLODAWER. 2002. Domain-swapped structure of a mutant of cyanovirin-N. Biochem. Biophys. Res. Commun. 294:184–190.

BOYD, M. R., K. R. GUSTAFSON, J. B. MCMAHON, R. H. SHOEMAKER, B. R. O'KEEFE, T. MORI, R. J. GULAKOWSKI, L. WU, M. I. RIVERA, C. M. LAURENCOT, M. J. CURRENS, J. H. CARDELLINA, R. W. BUCKHEIT, JR., P. L. NARA, L. K. PANNELL, R. C. SOWDER and L. E. HENDERSON. 1997. Discovery of

cyanovirin-N, a novel human immunodeficiency virus-inactivating protein that binds viral surface envelope glycoprotein gp120: potential applications to microbicide development. Antimicrob. Agents Chemother. 41:1521–1530.

COLLELUORIA, D. M., D. TIENA, F. KANGA, T. PAGLIEIA, R. KUSSA, T. MCCORMICKA, K. WATSONB, K. MCFADDENC, I. CHAIKENC, R. W. BUCKHEIT and J. W. ROMANOA. 2005. Expression, purification, and characterization of recombinant cyanovirin-N for vaginal anti-HIV microbicide development. Protein Express. Purif. 39:229–236.

DELANO, W. L. 2002. The PyMOL molecular graphics system. DeLano Scientific, Palo Alto, CA.http://www.pymol.org.

GASTEIGER, E., A. GATTIKER, C. HOOGLAND, I. IVANYI, R. D. APPEL and A. BAIROCH. 2003. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucl. Acids Res. 31:3784–3788.

GASTEIGER, E., C. HOOGLAND, A. GATTIKER, S. DUVAUD, M. R. WILKINS, R. D. APPEL and A. BAIROCH. 2005. Protein identification and analysis tools on the ExPASy server. The Proteomics Protocols Handbook. Humana Press, Totowa.

GEOURJON, C. and G. DELEAGE. 1995. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Comput. Appl. Biosci. 11:681–684.

GOPALASUBRAMANIAM, S. K., V. GARROCHO-VILLEGAS, G. B. RIVERA, N. PASTOR and R. ARREDONDO-PETER. 2008. Use of *in silico* (computer) methods to predict and analyze the tertiary structure of plant hemoglobins. Meth. Enzymology 436:393–410.

GOUET, P., E. COURCELLE, D. I. STUART and F. METOZ. 1999. ESPript: analysis of multiple sequence alignments in PostScript. Bioinformatics 15:305–308.

GUEX, N. and M. C. PEITSCH. 1997. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18:2714–2723.

GUSTAFSON, K. R., R. C. SOWDER, L. E. HENDERSON, J. H. CARDELLINA, J. B. MCMAHON, U. RAJAMANI, L. K. PANNELL and M. R. BOYD. 1997. Isolation, primary sequence determination, and disulfide bond structure of cyanovirin-N, an anti-HIV (human immunodeficiency virus) protein from the cyanobacterium *Nostoc ellipsosporum*. Biochem. Biophys. Res. Commun. 238:223–228.

HELLE, F., C. WYCHOWS, N. VU-DAC, K. R. GUSTAFSON, C. VOISSET and J. DUBUISSON. 2006. Cyanovirin-N inhibits hepatitis C virus entry by binding to envelope protein glycans. J. Biol. Chem. 281:25177–25183.

KOHARUDIN, L. M., A. R. VISCOMI, J. G. JEE, S. OTTONELLO and A. M. GRONENBORN. 2008. The evolutionarily conserved family of cyanovirin-N homologs: structures and carbohydrate specificity. Structure 16:570–584.

KOZAK, M. 1999. Initiation of translation in prokaryotes and eukaryotes. Gene 234:187–208.

LIU, Y. G. and R. F. WHITTIER. 1995. Thermal asymmetric interlaced PCR: automatable amplification and sequencing of insert end fragments from P1 and YAC clones for chromosome walking. Genomics 25:674–681.

MORI, T., R. H. SHOEMAKER, R. J. GULAKOWSKI, B. L. KREPPS, J. B. MCMAHON, K. R. GUSTAFSON, L. K. PANNELL and M. R. BOYD. 1997. Analysis of sequence requirements for biological activity of cyanovirin-N, a potent HIV (human immunodeficiency virus)-inactivating protein. Biochem. Biophys. Res. Commun. 238:218–222.

O'KEEFE, B. R., J. A. BEUTLER, J. H. CARDELLINA, R. J. GULAKOWSKI, B. L. KREPPS, J. B. MCMAHON, R. C. SOWDER, L. E. HENDERSON, L. K. PANNELL, S. A. POMPONI and M. R. BOYD. 1997. Isolation and characterization of niphatevirin, a human-immunodeficiency-virus-inhibitory glycoprotein from the marine sponge *Niphates erecta*. Eur. J. Biochem. 245:47–53.

O'KEEFE, B. R., D. F. SMEE, J. A. TURPIN, C. J. SAUCEDO, K. R. GUSTAFSON, T. MORI, D. BLAKESLEE, R. BUCKHEIT and M. R. BOYD. 2003. Potent anti-influenza activity of cyanovirin-N and interactions with viral hemagglutinin. Antimicrob. Agents Chemother. 47:2518–2525.

PAGE, R. D. 1996. TreeView: an application to display phylogenetic trees on personal computers. Comput. Appl. Biosci. 12:357–358.

PERCUDANI, R., B. MONTANINI and S. OTTONELLO. 2005. The anti-HIV cyanovirin-N domain is evolutionarily conserved and occurs as a protein module in eukaryotes. Proteins 60:670–678.

SAENZ-RIVERA, J., G. SARATH and R. ARREDONDO-PETER. 2004. Modeling the tertiary structure of a maize (*Zea mays* ssp. *mays*) non-symbiotic hemoglobin. Plant Physiol. Biochem. 42:891–897.

SAITOU, N. and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.

SCHWEDE, T., J. KOPP, N. GUEX and M. C. PEITSCH. 2003. SWISS-MODEL: an automated protein homology-modeling server. Nucl. Acids Res. 31:3381–3385.

SHAW, G. and R. KAMEN. 1986. A conserved AU sequence from the 3′ untranslated region of GM-CSF mRNA mediates selective mRNA degradation. Cell 46:659–667.

SU, Y. J., T. WANG, W. D. YANG, C. HANG and G. K. FAN. 1998. DNA extraction and RAPD analysis of *Podocarpus*. Acta Sci. Natur. Univ. Sunyatseni. 37:13–18.

THOMPSON, J. D., T. J. GIBSON, F. PLEWNIAK, F. JEANMOUGIN and D. G. HIGGINS. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucl. Acids Res. 25:4876–4882.

TRIGLIA, T., M. G. PETERSON and D. J. KEMP. 1988. A procedure for *in vitro* amplification of DNA segments that lie outside the boundaries of known sequences. Nucl. Acids Res. 16:8186.